DOCUMENT RESUME

ED 216 044 TM 820 208

AUTHOR Shah, Babubhai V.; And Others
TITLE Inferences About Regression Models. Contractor Report.
INSTITUTION Research Triangle Inst., Durham, N.C.
SPONS AGENCY National Center for Education Statistics (ED), Washington, DC.
REPORT NO NCES-82-210
PUB DATE Aug 81
CONTRACT OE-0-73-6666
NOTE 63p.

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Computer Programs; *Estimation (Mathematics); Graduate Surveys; High Schools; Least Squares Statistics; Mathematical Models; Multiple Regression Analysis; *Regression (Statistics); *Validity
IDENTIFIERS *Horvitz Thompson Estimator; Inference (Statistical); *National Longitudinal Study High School Class 1972

ABSTRACT

Aside from the theoretical issues involving the validity of inferences from surveys, the basic problem of producing unbiased estimates of regression parameters and estimates of the associated standard errors has been a particularly difficult issue in dealing with results from stratified multistage sample designs such as the one used in the National Longitudinal Study of the High School Class of 1972 (NLS). The purpose of this report is to review some appropriate available techniques that may be useful in applying regression models to the NLS data. The first section provides a framework for evaluation and an appraisal of some alternate approaches within this framework. A preferred approach (combining the Horvitz-Thompson estimator and taylorized deviation) is compared to an Ordinary Least Squares approach, through a simulation procedure using actual NLS data. The several results are summarized. Formulae underlying the preferred approach are provided separately in Appendixes A and B, and details of the development and use of a computer program to implement the approach are provided in Appendixes C and D. (Author/BW)

Contractor Report

Inferences About Regression Models

National Center for
Education Statistics

Inferences About Regression Models

Research Triangle Institute

Babubhai V. Shah
Mary Margaret Holt
Ralph E. Folsom

Andrew J. Kolstad
Project Officer
National Center for Education Statistics

August 1981

Prepared for the National Center for Education Statistics under contract OE-0-73-6666 with the U.S. Department of Education. Contractors undertaking such projects are encouraged to express freely their professional judgment. This report, therefore, does not necessarily represent positions on policies of the Government, and no official endorsement should be inferred. This report is released as received from the contractor.

NCES 82-210

ACKNOWLEDGMENTS

The authors wish to thank all those people who have contributed to this study and the writing of this report. Mr. Phillip Cooley wrote the generalized matrix inversion and other related routines that are incorporated in the program. Ms. Deborah Duncan wrote the SAS programs that generated the summary tables from the results of the simulation. Various comments and suggestions were provided throughout the entire work by Dr. D. G. Horvitz, Dr. Samuel Peng, and Dr. Jay Levinsohn. Special thanks also are due to Mr. William Fetters of NCES for providing many constructive suggestions, and to Dr. G. J. Burkheimer of RTI for editorial contributions.

Also, many people and organizations have contributed generously to the base-year and follow-up surveys, and their efforts are sincerely appreciated. We are especially grateful to the thousands of anonymous sample members who have participated in the surveys, without whose cooperation this continuing study would not have been possible.

CONTENTS

LIST OF TABLES

I. INTRODUCTION

The National Longitudinal Study (NLS) of the High School Class of 1972 is a large-scale sample survey sponsored by the National Center for Education Statistics (NCES). The sample design for this survey can be described as a deeply stratified two-stage design with 600 final strata. The original design called for 1,200 schools and 18 students per school (size permitting). A total of 1,069 schools and 16,683 students participated in the base-year survey, which was conducted by Educational Testing Service. An additional follow-up of nonrespondent schools, plus additional backup schools and augmentation of the sample for the first follow-up, increased the number of participating schools to 1,318 and the total student sample to 23,451. The numbers of respondents to the first, second, third, and fourth follow-up questionnaires, administered by the Research Triangle Institute (RTI), were 21,350, 20,872, 20,092, and 18,630, respectively.

As suggested above, a large amount of data has been collected for this study. The types of statistics required to address various research questions of interest range from simple descriptive totals and means to more complex analytic statistics, such as regression coefficients, but the problems of drawing valid and relevant inferences, which are common to all multistage sample surveys, must be addressed in analyzing the NLS data. For complex statistics, such as regression coefficients, there are no "pat" solutions to these problems; however, the need still exists for some good, even though "imperfect," techniques to approximate these statistics and their errors.

Aside from the various theoretical issues involving the validity of inferences from surveys, the basic problem of producing unbiased estimates of regression parameters and estimates of the associated standard errors has been a particularly thorny issue in dealing with results from stratified multistage sample designs such as the one used in the NLS. Most of the available statistical software packages [such as SPSS (Nie, et al., 1975), SAS (Barr, et al., 1976, 1977), BMDP (Dixon, 1975), or OSIRIS (Rattenbury and Eck, 1973; Institute for Social Research, 1973)] treat the sample as independent random observations, ignoring the sample design. This approach is convenient but theoretically inappropriate, since it does not account for unequal probabilities of selection or for effects of stratification and/or clustering. The application of sampling weights is possible through some software packages, allowing correct estimates

of regression coefficients, but appropriate error variance estimates typically are not produced. In fact, it is not possible to obtain explicit expressions for variance estimates of complex estimators such as regression coefficients within complex survey sample designs; however, various approximation procedures are available.

The purpose of this report is to review some appropriate available techniques that may be useful in applying regression models to the NLS data. The following section provides a framework for evaluation and an appraisal of some alternate approaches within this framework. In Section III, a preferred approach (combining the Horvitz-Thompson estimator and Taylorized deviation) is compared to an Ordinary Least Squares approach, through a simulation procedure using actual NLS data. The several results are summarized in Section IV. Formulae underlying the preferred approach are provided separately in Appendixes A and B, and details of the development and use of a computer program to implement the approach are provided in Appendixes C and D.

II. ASSESSMENT OF ALTERNATIVE TECHNIQUES

Survey research in the social sciences is often based on large complex samples, from which inferences are made regarding the population under study. The most common practice for drawing inferences about a univariate parameter is to assume $(\hat{\theta} - \theta)/s(\hat{\theta})$ has approximately the Gaussian or Student's t distribution, where the statistic $\hat{\theta}$ is an estimate of the parameter θ and $s(\hat{\theta})$ is an estimate of the standard error of $\hat{\theta}$. Similarly, for a multivariate parameter θ, represented as a row vector, inference may be based on a Hotelling's T^2 type statistic of the form $(\hat{\theta} - \theta)\{\hat{V}(\hat{\theta})\}^{-1}(\hat{\theta} - \theta)'$, which is assumed to have a chi-square or transformed F distribution in repeated samples, where $\hat{V}(\hat{\theta})$ is an estimate of the variance-covariance matrix of $\hat{\theta}$.

The justification for such an approach is based on the assumption that a generalized central limit theorem applies to large complex probability samples from finite populations. David (1938), Madow (1945), and Hájek (1960) have established such results for the mean of a simple random sample by letting the population size increase at the same rate as the sample size. For survey statisticians concerned with finite population inference, the regularity with which sampling distributions for properly standardized survey statistics can be expected to follow classical distributions continues to be one of the most important unanswered questions.

The problem is further complicated in the case of regression models in the specification of $\hat{\theta}$ and the standard error of $\hat{\theta}$, where θ is a vector of regression coefficients. A variety of models and interpretations have been suggested [e.g., Konijn (1962), Godambe and Thompson (1971), Royall (1971), Kish and Frankel (1974), Fuller (1974), and Folsom (1974)].

A. Overview and Notation[1]

For the NLS survey, schools were stratified by several characteristics to obtain $H = 600$ strata (Westat, 1972). Within each stratum, h, m_h (mostly 2) schools were selected at random from the total of M_h schools in the stratum. Within each school, (i.e, the ith school in the hth stratum), a random sample of n_{hi} (mostly 18) students from the total of N_{hi} students within school hi were selected for survey.

[1] Although the discussion in this and subsequent sections is sometimes specific to the NLS survey design, the results are generally applicable.

Within this context, estimates are required; for example, consider an estimate of the national total number of high school seniors who were in an academic curriculum. Let $\hat{X}_{hi}$ be the estimated number of students with academic curriculum in the (hi)th school, h = 1, 2, ..., H and i = 1, 2, ..., m_h.

If the sample of schools within each stratum is selected with equal probabilities and with replacement, then the estimated total, $\hat{T}$, and an unbiased estimate of its variance, $\hat{V}(\hat{T}_h)$, are

$$\hat{T} = \sum_{h=1}^{H} \frac{M_h}{m_h} \sum_{i=1}^{m_h} \hat{X}_{hi} = \sum_{h=1}^{H} \sum_{i=1}^{m_h} \hat{Y}_{hi} , \qquad (1.1)$$

and

$$\hat{V}(\hat{T}) = \sum_{h=1}^{H} m_h \sum_{i=1}^{m_h} (\hat{Y}_{hi} - \bar{\hat{Y}}_h)^2/(m_h - 1) , \qquad (1.2)$$

where $\hat{Y}_{hi} = \hat{X}_{hi} M_h/m_h$,

and $\bar{\hat{Y}}_h = \sum_{i=1}^{m_h} \hat{Y}_{hi}/m_h$.

If an approximate estimate of the size of schools (s_{hi} = size = total number of students in the (hi)th school) were known, then one could use a biased ratio estimator, $\hat{T}_1$, given by

$$\hat{T}_1 = \sum_{h=1}^{H} \sum_{i=1}^{m_h} \hat{Z}_{hi} , \qquad (1.3)$$

where $\hat{Z}_{hi} = s_{h+} \hat{X}_{hi}/m_h s_{hi}$,

and $s_{h+} = \sum_{i=1}^{M_h} s_{hi}$.

The estimator $\hat{T}_1$, which may be recognized as a Horvitz-Thompson (1952) estimator, is an unbiased estimator of the total, if the probability of selecting

the (hi)th school is $p_{hi} = s_{hi}/s_{h+}$ on each of m_h draws. An unbiased estimate of its variance is given as

$$\hat{V}(\hat{T}_1) = \sum_{h=1}^{H} m_h \sum_{i=1}^{m_h} (\hat{Z}_{hi} - \bar{Z}_h)^2/(m_h - 1), \qquad (1.4)$$

where $\bar{Z}_h = \sum_{i=1}^{m_h} \hat{Z}_{hi}/m_h$;

This appears to be the most common approach in many sample surveys. The absence of bias in the estimator $\hat{T}_1$ and the estimate of its variance $\hat{V}(\hat{T}_1)$ are established over repeated samples with the primary sampling units selected with unequal probabilities and with replacement.

While the prior discussion has been directed to providing an example of estimation within the NLS study, toward introducing notation, it also has illustrated the way in which the sample design or the conceptualization of all possible samples affects the variance of the estimator based on only one of the samples. The freedom of survey designers to define the sampling distributions has raised several fundamental issues regarding various statistical estimates. A full discussion of these issues is not within the scope of this report; however, the practical problem of selecting an appropriate estimator and an estimate of the variance of that estimator must still be addressed.

The estimator chosen for the current purpose is the Horvitz-Thompson estimator. For sampling with unequal probabilities, this estimator is used widely in practice and has been found to be an admissible estimate. The Horvitz-Thompson estimator is not the "best" estimator in all cases, but the same can be said of any other estimator. When probabilities of selection are based on prior information about size and the relationship of size to the characteristic of interest, the Horvitz-Thompson estimator is optimal or nearly optimal.

The choice of this estimator is not as arbitrary as it may appear (Shah, 1980); however, there are few practical rules to support the choice. The most common advice for selecting an estimator is to examine the data before deciding which estimator is optimal. An expert in survey design and theory may be able to reach such a decision because of past experience and knowledge. Other

researchers may need a catalog of alternate estimators and a set of rules that will enable them to select the optimal estimator. At present, no such guidelines exist except for such vague statements as, "If probabilities of selection have no relation to the characteristic to be measured then the simple mean would be better than the Horvitz-Thompson estimator." The survey practitioner obviously needs better guidelines for choosing estimators, but until such time as these rules become available, the survey practitioner probably will continue to use the Horvitz-Thompson estimator, which is optimal in most cases, even though it may be inefficient in a few situations.

A second important consideration, which is often neglected, is an estimate of the variance of the estimator. The estimator that one uses may or may not be optimal and may or may not be efficient, but it is imperative that some proper estimate of the variance (mean square error) of the estimator be computed from the data. The proper evaluation of mean square error assumes an additional dimension of importance when the estimator used is not unbiased. Guidelines for selecting from among available mean square error estimators also are not readily available.

In the case of estimates of error variance, there are additional considerations. For the example given above, the total is a simple linear function of the observations, and it is possible to derive explicit algebraic expressions for estimating variances of such linear functions. However, it is not possible to obtain such explicit expressions for variance estimates of complex estimators such as a regression coefficient or a correlation coefficient.[2] There are, however, various available approximation procedures; some such procedures are: (1) Taylorized deviations, (2) independent replications, (3) balanced repeated replications, and (4) Jackknife.

B. Application of the Central Limit Theorem

Assuming the estimate and variance for the total (1.1) and (1.2), let the vector $\hat{T}_h = (t_1, t_2, \ldots, t_k)_h$ represent the totals of k variables $(x_1, x_2, \ldots, x_k)$ for the hth stratum. An estimator of total $\hat{T}_h$ and its variance-covariance matrix $\hat{V}(\hat{T}_h)$ can be obtained using formulae similar to (1.3) and (1.4). Further, let the vector $\hat{T}$ denote the sum of the vectors $\hat{T}_h$. Since the sampling

[2] It should be noted that this difficulty with complex statistics is common to all branches of statistics and is not a distinctive feature of sample surveys.

within one stratum is independent of sampling within another, it follows that

$$\hat{T} = \sum_{h=1}^{H} \hat{T}_h \quad , \tag{1.5}$$

and an estimate of the variance-covariance matrix of $\hat{T}$ is

$$\hat{V}(\hat{T}) = \sum_{h=1}^{H} \hat{V}(\hat{T}_h) \quad . \tag{1.6}$$

If a large number of strata[3] are involved and it is assumed that the first two moments of the distributions of $\hat{T}_h$ (h = 1, 2, ..., H) satisfy certain convergence properties (e.g., Lindberg conditions), a general form of the central limit theorem would apply (Feller, 1966); hence, the limiting distribution of $\hat{T}$ would be multivariate normal.

If one is interested in estimating the variance of a statistic, $\hat{\theta}$, which is a nonlinear function of $\hat{T}$, then the approximate normality of $\hat{T}$ is not necessarily useful in estimating $\hat{V}(\hat{\theta})$. Examples of such nonlinear functions are:

$$\hat{\theta}_1 = \frac{\Sigma w_h x_h}{\Sigma w_h} \quad ,$$

and

$$\hat{\theta}_2 = \frac{(\Sigma w_h x_h y_h - \Sigma w_h x_h \Sigma w_h y_h / \Sigma w_h)}{\sqrt{\{\Sigma w_h y_h^2 - (\Sigma w_h y_h)^2 / \Sigma w_h\} \{\Sigma w_h x_h^2 - (\Sigma w_h x_h)^2 / \Sigma w_h\}}} \quad .$$

The statistics $\hat{\theta}_1$ and $\hat{\theta}_2$ can be readily recognized as the weighted mean of x and the weighted correlation between x and y, respectively, where w_h represents the weight.

---

[3] If sampling of Primary Sampling Units (PSUs) is with replacement, the same arguments can be made at PSU levels.

C. Variance Approximation Procedures

As stated previously, four relatively common approaches to appropriate variance estimation are (1) Taylorized deviations, (2) independent replications, (3) pseudo-replications, and (4) Jackknife. Brief descriptions of each of these techniques, their assumptions, and their strengths and weaknesses are provided in this section.

1. Taylorized Deviations

A classical solution to the estimation problem has been to express the statistic $\hat{\theta}$ as a polynomial in $(t_1, t_2, \ldots, t_k)$ elements of the vector T, using the Taylor Series expansion. The approximate variance of $\hat{\theta}$ can then be obtained by using only the linear terms of this expansion (see Kendall and Stuart, 1973).

If $(\delta\theta/\delta T)$ is a row vector of derivatives, $\{\delta\theta/\delta t_1, \delta\theta/\delta t_2, \ldots, \delta\theta/\delta t_k\}$, then the approximate variance of $\hat{\theta}$ is estimated by

$$\hat{V}(\hat{\theta}) = (\hat{\delta\theta}/\hat{\delta T})\hat{V}(T)(\hat{\delta\theta}/\hat{\delta T})' ,$$

which can be further expanded as

$$\hat{V}(\hat{\theta}) = \sum_{h=1}^{H} (\hat{\delta\theta}/\hat{\delta T_h})\hat{V}(\hat{T_h})(\hat{\delta\theta}/\hat{\delta T_h})' .$$

For large values of H, it is assumed that the distribution of $\hat{\theta}$ will be approximately normal with variance $\hat{V}(\hat{\theta})$. Such expansion for ratio estimates is presented in most textbooks on sample surveys. The first-order Taylor Series expansion for regression coefficients has been derived by Folsom (1974) and Fuller (1974). Woodruff (1971) has presented an algorithm for obtaining a first-order Taylor Series approximation to compute the variance of any complex statistic. Programs for Taylorized deviations are available from Hidiroglou and Fuller (1975), Holt (1977), Kish et al. (1972), Shah (1974), and Woodruff and Causey (1976).

2. Independent Replications

The most straightforward way to avoid assumptions would be to draw several independent samples from the same population and, thus, to obtain several independent estimates of the same statistic $\hat{\theta}$ (i.e., $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_r$).

The mean estimate, $\hat{\theta}$, would be

$$\bar{\theta} = \sum_{i=1}^{r} \hat{\theta}_i / r$$

and an estimate of the variance, $\hat{V}(\bar{\theta})$, is given by

$$\hat{V}(\bar{\theta}) = \sum_{i=1}^{r} (\hat{\theta}_i - \bar{\theta})^2 / r(r-1) .$$

In practice, however, one would like to compute $\hat{\theta}$ using data from all the samples, and for complex statistics $\hat{\theta}$ will not necessarily be equal to $\bar{\theta}$. It is then necessary to assume that $\hat{V}(\bar{\theta})$ is approximately equal to $\hat{V}(\hat{\theta})$.

A practical problem exists with this technique in that it places severe restrictions on the sample design, since each independent sample is much smaller than the "total sample" feasible with limited resources. Further, resources may similarly constrain the number of independent replications (samples) to be small; consequently, the estimate of the variance would have few degrees of freedom and would tend to be unstable. Additionally, in the case of multivariate analysis where θ is a vector of dimension P, if $P > r$, then the estimated variance-covariance matrix $\hat{V}(\hat{\theta})$ will be singular.

3. Pseudo-Replications

An ingenious but simple approach was suggested by McCarthy (1966) for designs with exactly two primary sampling units (PSUs) per stratum. A random half of the sample is defined by randomly selecting one of the PSUs in each stratum; the half sample and its complement are assumed to be "approximately" independent samples. Thus, an estimate of the variance with one degree of freedom can be computed using two half samples. Of course, it is necessary to assume that the variance of the statistic based on the total sample is approximately half that of the estimate based on half replicates. Since there are 2^H possible half samples, many pairs of half samples can be selected. In practice, about 40 to 100 pairs of half samples are selected to provide reasonable estimates of the variances.

The determination of the approximate degrees of freedom for the estimated variance remains an unanswered question. The practical approach is to assume degrees of freedom equal to the number of strata or the number of pairs of

half replicates, whichever is smaller. If both of these are large (i.e., greater than 30), then, in practice, the actual value is irrelevant since the t or F distributions can be approximated by the normal or χ^2 distributions, respectively.

4. Jackknife

The "Jackknife" approach originally suggested by Quenoille (1956), and so named by Tukey (1958), is an intuitive approach to computing variances. A definition of "Jackknife" for a multistage survey design in which all stages are random is presented by Folsom et al. (1971). Kish and Frankel (1974) have suggested an approach for a stratified sample with two PSUs per stratum; however, no general definition is available for a stratified multistage sample.

D. Additional Considerations

1. Computation

Most of the widely used statistical packages (e.g., SPSS, BMDP, OSIRIS, SAS) do not routinely provide for computing proper variances of a weighted statistic from a multistage sample survey. Except at institutions with large statistical and computational resources, the computation of such standard errors frequently is not attempted.

Frequent complaints are that the cost of computing variances is excessive and that standard software for the computation is not available (the cost of special purpose programming being prohibitively expensive). For example, the cost of computing the variance of a weighted mean may be 10 to 50 times that of computing the mean. While this may be the case for some techniques or programs, RTI's experience in using the Taylorized deviation approach is that the total cost of computing variances is only about twice that of computing only the mean. Moreover, several general-purpose programs have become available recently (see subsection II.C.1, above).

2. Estimating Variance Components

Many surveys are conducted periodically, and there is a need for evaluation of survey designs used with a view to possible improvements in subsequently designing similar surveys. To make decisions about such designs, there is a need to estimate contributions to the variance of a statistic from various elements of the overall design such as stratum, PSU, and individual; in other words, estimation of variance components is required. Of the techniques discussed above, Taylorized deviation is the only one that permits estimation

of variance components (see Shah et al., 1973; Moore et al., 1974). Since the estimator is expressed as a sum of random variables, the variance components of $\hat{\theta}$ can be estimated in the same manner as that of T.

E. Comparison of Techniques

To compare the techniques, the following criteria are used:

1) validity or number of assumptions required,
2) restrictions on sample design,
3) computational problems for large data sets, and
4) flexibility of applications.

A summary of the comparison is presented in Table 1. From the comparison, the Taylorized deviation approach appears to be best, if one is willing to accept applicability of the central limit theorem. Furthermore, if one needs to evaluate components of variance, then Taylorized deviation is the only approach. If there are only two PSUs per stratum in the design, pseudo-replications would be appropriate.[4] The Jackknife approach should be considered only in the rare case of a complex design and for a statistic for which it is not possible to evaluate derivatives. The independent replications approach will be suitable only if the sample is designed appropriately.

F. Conclusions

The recommendation supported by the discussion in this section is that for most nontrivial survey designs, the Horvitz-Thompson estimator and a Taylorized deviations approach are typically the most appropriate and practical techniques for computing parameter estimates and associated estimates of the variance, including estimates of regression coefficients. The choice of the Horvitz-Thompson estimator is based partially on intuitive grounds but is also supported theoretically (Shah, 1980). The choice of Taylorized deviations was made for the following reasons: (1) applicability to all designs and statistics; (2) applicability to large samples; (3) economy and computational feasibility; and (4) capacity for estimating variance components.

The assumption underlying the Taylorized deviations approach is asymptotic normality. The assumption of approximate normality is in use in other contexts, and some rules of thumb have been developed (e.g., a binomial distribution is

[4] Although the original NLS survey design had two schools per stratum, the ultimate design had several strata with three or four schools.

Table 1.--Summary of comparative evaluation

Technique	Criteria			
	Assumptions	Restrictions on sample design	Computational problems	Flexibility
Independent replications	Minimal	Severe	Simple	--
Psuedo-replications	Independence of complementary half replicates	Two PSUs per stratum	Significant	--
Taylorized deviations	General central limit theorem	None	Not difficult	Can be used for variance components
Jackknife	Intuition	None	Greater than Taylorized deviation	May be useful for some designs

approximately normal if npq is greater than 10). There is an obvious need for developing such simple rules of thumb for statistics resulting from survey samples; however, until more information is available, the suggested approach is Taylorized deviations using any of the available programs (Hidiroglou et al., 1975; Holt, 1977; Shah, 1974; or Woodruff and Causey, 1976). In practice, one should consider certain transformations of statistics that rapidly converge to normality; as an example, if r is the sample correlation, then evaluation of the variance of $\text{Tanh}^{-1}(r)$ may be more appropriate.

A development of the Taylorized deviations approach for regression coefficients is provided in Appendixes A and B, for the interested reader. A flexible and easily used computer program applying Taylorized deviations to the computation of regression coefficients and their standard errors for data arising from multistage samples is described in Appendixes C and D. This program is available from the senior author of this report.

III. EMPIRICAL TESTS OF TWO REGRESSION APPROACHES

Previous discussion has indicated the theoretical superiority, under assumptions of approximate normality, of a combined Horvitz-Thompson estimator and a Taylorized deviation variance approximation approach to the investigation of regression models with data arising from complex survey samples. Nonetheless, there is need for some empirical evidence of the verity of the approach; consequently, a simulation procedure was undertaken, using the NLS data base. The study involved drawing a large number of random samples from a finite population and then deriving pertinent statistics from these samples, to evaluate the distribution of the regression coefficients and that of the approximate F values computed by Taylorized deviations.

The simulation also allowed a natural vehicle for evaluation of other, less appropriate, approaches to regression analysis as compared to the suggested approach. One of the most widely used approaches to regression analysis ignores the sample design and addresses the data as though they arose from a simple random sample. This approach, using Ordinary Least Squares (OLS) criteria, owes much of its popularity to the facts that it is better known than the more appropriate techniques and that it is easily applied through all of the widely used statistical analysis packages. Nathan and Holt (1980) have demonstrated that in most cases the regression coefficients computed by applying OLS solutions to data collected from complex survey designs will be biased, although an exception occurs for epsem designs. Moreover, they showed that, under these conditions, the OLS variance estimator is consistently biased even in those cases for which the OLS regression coefficient estimates themselves are unbiased. While the proper application of sampling weights, within some standard statistical packages, can produce unbiased estimates of the regression coefficients, the weighted variance estimate produced by most packages remains biased. Moreover, RTI's experience with this latter approach suggests that resulting variances show considerably greater bias than those obtained through OLS. For these reasons, OLS was chosen as the comparison approach to Taylor Series Linearization (TSL).

A. Method

The NLS base-year sample was taken as the finite population for this simulation. The original sample design consisted of 600 strata, with 2

schools selected per stratum. Within each school, an equal probability sample of roughly 18 seniors was selected. For the simulation, 84 major strata were formed by combining similar strata, each containing at least 10 schools. So as not to confound results of the simulation study with problems of missing data the finite population was defined to exclude students with missing data elements for any of the variables used in the regression models. Consequently, the study population contained 935 schools and 10,657 students.

The simulation consisted of selecting 1,000 random samples from the defined population. Each sample was selected in two stages. Within each stratum, two schools were selected without replacement with probabilities proportional to estimated senior class enrollments; Durbin's (1967) method was used for these selections. From each of the schools in each sample, five responding students were selected with equal probabilities and without replacement; thus, each resulting sample consisted of 840 students. OLS and TSL values of regression coefficients and their associated variances, covariances, and F values were computed for each of the 1,000 samples and for 4 regression equations.

Two basic regression models were selected for the NLS simulation study, and within each model two related criterion variables were used to indicate the type of postsecondary education being received by individuals in the fall of 1974. This resulted in four regression equations for evaluation, although the two criterion variables for each model were similar (both related to type of postsecondary entry, but one was a dichotomization of the other). The predictor variables represented characteristics of the high school seniors prior to graduation in 1972. The two basic regression models are written symbolically below.

Model 1: INC (or TYP) = INT + SEX + SES + GRADES + GOALS.

Model 2: INC (or TYP) = INT + SEX + SES + ABIL + RACE*PROG.

The variables used in the two models are defined below.

INC = 1 if the individual had enrolled in some type of education after high school;
0 otherwise.

TYP = type of college enrollment, scaled as follows:
4 if 4-year college;
3 if 2-year college;
2 if any other regular or vocational college;
1 if not enrolled in any college.

INT = the model intercept.

SEX = 0 if female;
1 if male.

SES = composite socioeconomic-educational status derived from several base-year questionnaire items (see Dunteman, et al., 1974).

GRADES = self-reported, overall, high-school grade range (8 levels).

GOALS = a quantitative measure of educational or other aspirations derived from base-year questionnaire response (see Dunteman, et al., 1974).

ABIL = an ability score based on test items administered during the base-year (see Dunteman, et al., 1974).

RACE*PROG = indicator variables for the joint contribution of race/ethnicity and high school program including their interaction, where

$R*P_1$ = 1 if the high school program was academic and race/ethnicity was majority white,
0 otherwise;

$R*P_2$ = 1 if the high school program was academic and race/ethnicity was any minority,
0 otherwise;

$R*P_3$ = 1 if the high school program was nonacademic and the race/ethnicity was majority white,
0 otherwise;

$R*P_4$ = 1 if the high school program was nonacademic and the race/ethnicity was any minority,
0 otherwise.

The regression models were evaluated using both OLS and TSL, as applied through procedure SURREGR described in Appendixes C and D. In full model (ALL) hypotheses, the intercept was excluded. Also, the RACE*PROG hypothesis of model 2 was reduced to rank 3 by eliminating the $R*P_4$ variable. The variance of the regression coefficients, the mean difference from the corresponding population value, and the standard error of the mean were computed over the

1,000 samples in order to evaluate the possible bias in estimates of the coefficients. Means of the estimated variances were also computed. Additionally, the values of the F statistic for testing the hypothesis that the regression coefficients were equal to known population values were computed. Since the null hypothesis is true, the observed values should resemble the theoretical F distribution. The actual numbers of observed F values falling below various percentile points of appropriate F distributions were tabulated for this comparison.

B. Results

The discussion in the previous section leads to three predictions which should be observable from the results of simulation, if the estimators and their variances are unbiased.

1) The expected value of the difference of each regression coefficient from its true value over all samples should be approximately 0; therefore, the mean value over all samples of a regression coefficient should fall within the interval defined by the true value ± 3 times its standard deviation.

2) The expected value of the variance of a regression coefficient which was computed by the Taylorized deviation method should be approximately equal to the variance of that regression coefficient over all samples.

3) The percentage over all samples of the statistically significant F values for testing a hypothesis about the difference of computed coefficients from known population values should be approximately equal to the nominal significance level.

Summary statistics to check the validity of predictions 1 and 2 are presented in Tables 2 and 3, for the TSL and OLS approaches, respectively. The first two columns of each table define the four regression equations examined (i.e., the criterion variable and predictor variables of the two basic models, respectively). The entries in column 3 give, for each predictor variable, the average (over the 1,000 samples) of the difference between the estimated regression coefficients and the actual population value of that coefficient. The estimated standard errors of these mean differences are given in column 4. The variances of the estimated regression coefficients over the 1,000 samples are provided in column 5, and the averages over the

Table 2.--Statistics describing the distribution of the estimated regression coefficients over 1,000 samples from a finite population using Taylor Series Linearization

Criterion variable	Predictor variable	Mean difference from population value	Standard error of the mean	Variance of the computed coefficients	Mean of the computed variances
INC	INT	0.02426	0.00254	0.00643	0.00747
	SEX	0.01225	0.00110	0.00121	0.00132
	SES	-0.00296	0.00078	0.00060	0.00065
	GRADES	-0.00303	0.00040	0.00016	0.00017
	GOAL	-0.00493	0.00046	0.00021	0.00023
TYP	INT	-0.06369	0.00644	0.04150	0.04563
	SEX	0.02666	0.00276	0.00764	0.00830
	SES	-0.00714	0.00200	0.00399	0.00415
	GRADES	-0.00682	0.00102	0.00105	0.00110
	GOAL	-0.01487	0.00118	0.00139	0.00152
INC	INT	-0.02038	0.00456	0.02078	0.02375
	SEX	0.00684	0.00113	0.00129	0.00134
	SES	0.00973	0.00086	0.00074	0.00083
	ABIL	0.00016	0.00002	0.00001	0.00001
	RP1	-0.01228	0.00274	0.00750	0.00831
	RP2	-0.00974	0.00311	0.00967	0.01030
	RP3	-0.00756	0.00251	0.00628	0.00642
TYP	INT	-0.03576	0.01164	0.13541	0.15359
	SEX	0.01376	0.00292	0.00851	0.00859
	SES	-0.02292	0.00220	0.00484	0.00543
	ABIL	0.00030	0.00006	0.00000	0.00000
	RP1	-0.03221	0.00698	0.04865	0.05542
	RP2	-0.03645	0.00805	0.06475	0.07237
	RP3	-0.01750	0.00629	0.03951	0.04111

Table 3.--Statistics describing the distribution of the estimated regression coefficients over 1,000 samples from a finite population using Ordinary Least Squares

Criterion variable	Predictor variable	Mean difference from population value	Standard error of the mean	Variance of the computed coefficients	Mean of the computed variances
INC	INT	-0.025980	0.001963	0.003855	0.003923
	SEX	-0.004871	0.000834	0.000695	0.000722
	SES	-0.004038	0.000607	0.000368	0.000383
	GRADES	0.000728	0.000320	0.000102	0.000104
	GOAL	0.003870	0.000349	0.000122	0.000115
TYP	INT	-0.029836	0.005044	0.025448	0.025306
	SEX	-0.015392	0.002112	0.004460	0.004658
	SES	-0.025277	0.001603	0.002572	0.002473
	GRADES	0.003476	0.000822	0.000676	0.000671
	GOAL	0.004050	0.000908	0.000824	0.000745
INC	INT	-0.018672	0.003597	0.012941	0.013219
	SEX	-0.009746	0.000910	0.000828	0.000828
	SES	-0.007226	0.000678	0.000460	0.000519
	ABIL	0.000145	0.000017	0.000000	0.000000
	RP1	-0.012951	0.002193	0.004811	0.004562
	RP2	-0.013945	0.002629	0.006914	0.007318
	RP3	-0.014713	0.001963	0.003855	0.003617
TYP	INT	-0.003198	0.009129	0.083348	0.087252
	SEX	-0.026208	0.002335	0.005452	0.005469
	SES	-0.036864	0.001741	0.003034	0.003429
	ABIL	0.000328	0.000044	0.000002	0.000002
	RP1	-0.053305	0.005632	0.031729	0.030115
	RP2	-0.078817	0.006933	0.048073	0.048297
	RP3	-0.044036	0.004948	0.024485	0.023876

1,000 samples of the variance estimates computed for each sample are given in column 6.

Prediction 1 can be examined from the entries in column 3 and 4 of Tables 2 and 3. The mean differences of computed and actual regression coefficient values are clustered near 0, ranging from -.064 to .027 for the Horvitz-Thompson and Taylorized deviation approach and from -.079 to .004 for the OLS approach. With few exceptions, however, the confidence intervals of three standard errors about these means did not include the value of 0, which implies some bias in estimating the regression coefficients by both of the approaches.

Prediction 2 can be examined from the entries in columns 5 and 6 of Tables 2 and 3. The TSL variance for each sample was computed according to equation (A.32), as provided in Appendix A. The average, over samples, of the TSL variance estimates is quite comparable to the actual variance, over the 1,000 samples, of the computed regression coefficients (Table 2). Similar results are also observable for the analogous OLS statistics (Table 3).

Summary statistics to check the validity of prediction 3 are provided in Tables 4 and 5. These tables indicate for regression models 1 and 2, respectively, a comparison of the upper tail of the appropriate theoretical F distribution to the empirical distribution of F values computed for each hypothesis in each of the 1,000 simulations. Within each of these tables, results are presented separately for each of the criterion variables considered in the particular model and for TSL and OLS approaches.

The TSL solutions appear to give good approximations for both models and for both criterion variables. Using an average of the empirical distributions over the various hypothesis tests within model and criterion variable, the TSL solutions can be seen to approximate the theoretical percentage points quite well. With one exception, such averages differ from nominal values by no more than one-half of a percentage point, and all such differences are in a conservative direction (i.e., suggest the null hypothesis would have been rejected less frequently than suggested by the nominal significance level).

In general, the OLS solutions also provide good approximations to the theoretical F distributions. The average of empirical distributions suggests that OLS solutions tend to err in a nonconservative direction and that the error is greater in modeling the criterion variable TYP. Even though the average differences from the theoretical distribution are still relatively small in an absolute sense (at most 3.5 percentage points), a question is

Table 4--The number of sample F values (out of 1,000) which fell below specified percentiles of the appropriate F distribution for Model 1

Analysis type	Criterion variable	Source	Hypothesis degrees of freedom	F distribution percentile				
				75	90	95	97.5	99
Taylor Series Linear-ization	INC.	ALL	4	751	893	950	978	993
		SEX	1	745	896	947	980	990
		SES	1	762	917	953	974	992
		GRADES	1	760	904	954	979	992
		GOALS	1	750	911	962	986	993
	TYP	ALL	4	755	898	960	981	992
		SEX	1	751	907	956	977	992
		SES	1	754	916	957	977	991
		GRADES	1	755	903	941	975	987
		GOALS	1	736	896	950	976	997
Ordinary Least Squares	INC	ALL	4	745	902	955	974	989
		SEX	1	751	895	952	975	993
		SES	1	745	902	949	974	989
		GRADES	1	758	906	951	973	987
		GOALS	1	716	864	931	953	978
	TYP	ALL	4	732	868	934	973	988
		SEX	1	744	899	947	969	991
		SES	1	674	846	920	956	977
		GRADES	1	755	896	943	967	987
		GOALS	1	737	871	930	962	980

Table 5.--The number of sample F values (out of 1,000) which fell below specified percentiles of the appropriate F distribution for Model 2

Analysis type	Criterion variable	Source	Hypothesis degrees of freedom	F distribution percentile				
				75	90	95	97.5	99
Taylor Series Linear-ization	INC	ALL	6	748	892	944	970	986
		SEX	1	736	905	952	975	992
		SES	1	763	894	954	971	990
		ABIL	1	777	919	965	988	995
		RACE*PROG	3	749	900	944	970	986
	TYP	ALL	6	755	889	953	972	987
		SEX	1	756	908	953	973	987
		SES	1	744	893	940	975	992
		ABIL	1	775	919	967	982	997
		RACE*PROG	3	769	899	947	972	987
Ordinary Least Squares	INC	ALL	6	776	893	944	971	987
		SEX	1	707	887	941	971	987
		SES	1	749	906	955	977	988
		ABIL	1	759	900	942	976	990
		RACE*PROG	3	757	892	947	974	986
	TYP	ALL	6	709	870	922	956	978
		SEX	1	713	877	936	971	988
		SES	1	684	841	913	957	986
		ABIL	1	744	902	953	977	991
		RACE*PROG	3	732	881	933	968	989

raised as to the extent to which the applicability of OLS approximations is situational (see particularly the poor fit for SES, in OLS solution for TYP).

C. Additional Simulations

Although not specific to the NLS data, similar simulations (Shah, et al., 1977) compared TSL and OLS under a wider range of sampling situations with different populations defined from the Health and Nutrition Examination Survey (cf., Public Health Service, 1973). In general, these results further support the contention that the agreement of OLS solutions with the theoretical F distribution is situational. Using 24 strata and selecting, first, 2 of 12 PSUs per stratum and, second, cluster sizes of 10 from each PSU, regression models similar to those used in the NLS simulations were employed. As with the NLS simulations, TSL solutions, in general, gave only marginally better results than OLS; however, the performance of the OLS statistic was again generally nonconservative and was better for continuous variables and for the case of greater between-PSU homogeneity (which reduces clustering effects). The performance of the TSL statistic was again relatively stable over all conditions.

In the special case of the application of regression models to solutions for domain means, the overall superiority of the TSL statistic was quite pronounced.[5] Under these conditions, not only did the OLS statistic generally show considerably less congruence than TSL to the theoretical F distribution, but also the congruence of the OLS statistic varied dramatically with different cluster sizes, different strata definitions, different between-PSU heterogeneity, and different models. With the exception of the prediction of domain means for race/ethnicity with a small number (8) of defined strata, the TSL F statistic was quite consonant with the theoretical F distribution and otherwise showed little variability from situation to situation.

D. Concluding Remarks

It should be recognized that the results of the several simulations offer only support and not definitive proof of the general applicability of the TSL

[5] Although this is a fairly atypical application of regression modeling, it does demonstrate the general applicability of the TSL approach over a wide variety of situations and the lack of such applicability for the OLS approach. For a technique of adjusting standard errors for domain means computed from NLS data, see Williams, 1978.

approach. No amount of empirical data can conclusively prove that any statistic provides valid inference in general. The simulations with NLS data were quite limited, restricted to two relatively simple (and related) regression models crossed with two (related) criterion variables. The consideration of results from the additional simulations provides a somewhat broader but still limited base for conclusions. The additional simulations examined only three regression equations (for each of two defined sampling frames) of a form similar to those examined in the NLS simulations. The additional results also included 64 simulations of the special case of applying regression modeling to computation of domain means. While these latter results are certainly germane to the general applicability of a statistical procedure, they do not directly address more conventional regression approaches.

Also, it should be recognized that the actual NLS data base differs from the NLS simulation in some important ways. While the NLS simulation results were based on a cluster size of 5, the actual NLS cluster size is 17. Other things being equal, increased cluster size tends to increase the variability of statistics. As an example, for a simple statistic (note that a regression coefficient is not a simple statistic), the impact of cluster size, m, on the variance, σ^2, can be indicated by the straightforward equation

$$\sigma_c^2 = \{1 + (m-1)\rho\}\sigma^2 ,$$

where σ_c^2 is the variance including the clustering effect and ρ is the intra-class correlation. Thus, the clustering effect for simple statistics, $(m-1)\rho$, would be expected to be four times larger with the actual NLS data than in the simulations (i.e., $(17-1)/(5-1) = 4$), with equivalent values of ρ. Further, even if ρ were as small as .02 an increase in actual variance of 32 percent (i.e., 16(.02)) over that of random sampling would be expected for simple statistics, other things being equal.

Finally, it should be recognized that the simulations were conducted under more or less ideal conditions in regard to comparability of sample weights, a situation that should theoretically favor OLS, which assumes equal sampling weights. While no marked disparities of sampling weights exist for the NLS data, some differences in weights have been introduced as a result of oversampling and adjustments for nonresponse (recall that the simulations did not address the problem of nonresponse). Under such conditions, bias in the OLS estimators may be expected to increase.

With an understanding of these additional considerations as well as the limitations of the results and drawing from results of both simulation studies, the general findings support the following conclusions.

1) TSL, though not perfect, produces "good" conservative inferences for regression coefficients (i.e., the probability of rejecting the null hypothesis, when true, is smaller than the nominal value) when the number of strata is moderately large (greater than 20).

2) In some situations, the performance of TSL is less satisfactory when the number of strata is small (less than 10).

3) OLS produces nonconservative inferences for regression coefficients (i.e., the probability of rejecting the null hypothesis, when true, is larger than the nominal value); however, in some situations the extent of nonconservatism is negligible.

4) While OLS compares favorably to TSL in the specific typical regression models considered, there are indications that the extent of acceptability of the technique is situational.

5) In the special case of domain means, the results of OLS are generally poor. Moreover, in this situation, the performance of OLS deteriorates considerably when the cluster size is increased.

IV. SUMMARY AND CONCLUSIONS

The sample design for NLS is a deeply stratified two-stage design with 600 final strata. Although the original design called for 1,200 schools (2 per stratum) and approximately 21,600 students (18 per school), the final sample, as defined from various sample additions, included 1,318 schools and 23,451 students. The information collected during the NLS base year and in subsequent follow-up surveys represents a rich data source for addressing questions regarding the educational and occupational development of high school graduates. The types of statistics used to address these questions may vary from simple totals to ratio and regression estimators; however, the problems of drawing valid and relevant inference that are common to all multi-stage sample surveys must be faced. The "perfect" answers to drawing inferences for complex statistics from survey data may not be readily available, but an applied scientist needs some good, though imperfect, techniques to provide approximate quantitative measures for the errors in the estimates.

This report has reviewed available theories and has suggested a technique that will be useful in analyzing NLS data with respect to regression models. For drawing inferences, it is imperative that some estimate of the variance (mean square error) of the estimator be computed from the data. For a simple linear function (such as a total or mean) of the observations, it is possible to derive explicit algebraic expressions for estimating variances; however, it is not possible to obtain such explicit expressions for variance estimates of complex estimators such as regression coefficients. The approximation procedures considered were: (1) Taylorized deviations, (2) independent replications, (3) balanced repeated replications, and (4) Jackknife. The Taylorized deviations approach is preferred for the following reasons: (1) it is applicable to all designs and statistics; (2) it provides "good" answers for "large" samples; (3) it is economically and computationally feasible; and (4) it alone provides for estimation of variance components.

Since the applicability of the Taylorized deviations approach is based on asymptotic theory, its performance was evaluated empirically through simulation, using NLS data. Additional simulations using another large data set were also considered. Simulations were carried out using both Taylor Series Linearization (TSL), as defined in Appendix A, and Ordinary Least Squares (OLS). Aside from potential violations of assumptions of the relatively robust regression

model, OLS is obviously inappropriate for drawing inferences from complex samples, assuming as it does simple random sampling; nonetheless, the technique was considered because it is so widely known and used (even when theoretically inappropriate) and is so easily applied through the more widely used statistical packages.

The simulations, though limited in scope, do suggest that TSL performs extremely well in a large variety of situations. With a small (i.e., less than 10) number of strata, there is some deterioration in its performance in some cases, but there is a dramatic improvement in performance with more than 20 strata. Both TSL and OLS show some bias in the estimation of regression coefficients; however, errors in inference using TSL are generally conservative, while the OLS approach generally yields nonconservative results (i.e., statistical tests are likely to reject the null hypothesis more frequently than they should). In some typical regression situations, however, the nonconservatism of OLS is negligible, and the approach performs quite well. Nonetheless, there are clear indications that the extent to which OLS solutions approximate the theoretical F distribution is situational. OLS performs particularly poorly in the specific case in which a regression model is applied to the estimation of domain means.

The various findings do suggest some practical recommendations to those who wish to use regression models in analysis of NLS data. In making general recommendations for use of a statistical methodology, even for a specific survey or specific hypothesis testing, the performance of the methodology in a broad variety of situations is relevant. Thus, if a methodology is successful on one or two hypotheses for a specific survey, there is no logical justification that it will perform well for all similar hypotheses, even with the same data. On the other hand, a methodology that is successful for several different hypotheses and different data sets may be expected to perform reasonably well in most situations. Moreover, a fairly general rule in applied statistics is that, given equality in other areas, recommended statistical methodologies which have potential for erring should err in a conservative direction.

Under these guidelines, the TSL procedure can be recommended for the NLS data. In fact, the transformed Hotelling's T^2 type statistic, using the TSL variance-covariance matrix, provides fairly robust multivariate inferences about regression coefficients with a moderately large number of strata (i.e., 24 or more). Although standard software for use of TSL is not widely available,

such software does exist (see Section II.C.1). The procedure SURREGR described in Appendixes C and D can generally be supported on a system supporting SAS; this procedure is available from the senior author of this report.

Although OLS yielded good results for some regression models in the simulations, it cannot be recommended for general use on the NLS data base. Not only is the OLS procedure logically poor when compared to TSL (OLS results are necessarily biased when applied to the NLS design--see Nathan and Holt, 1980), but also it is nonconservative. The potential user of the NLS data base may be tempted to use an OLS regression approach on the basis of the fact that OLS appeared to perform reasonably well in most simulations involving typical regression models. Such a decision would involve, of course, an element of risk, since there is an indication that OLS does not perform equally well for all models or designs. Moreover, the actual NLS data base differs from the NLS simulation in some important ways. Specifically, the actual NLS data are based on larger cluster sizes and contain more disparate sample weights.

Even though the OLS procedure cannot be recommended for general use with NLS data, it should be noted that the principal purpose of this research was not to examine the robustness of OLS. Additional research obviously is needed to determine the conditions under which OLS regression solutions might acceptably approximate those of more appropriate approaches. Further, the recommendations provided above have addressed the situation of drawing inferences from a sample (i.e., estimating population parameters); however, many regression studies are not directed to this end. For such other uses of regression with the NLS data (e.g., sample-specific modeling, exploratory studies), the use of OLS may be more appropriate, but such uses also are beyond the scope of this study. In such cases, however, the potential user must recognize the clear distinction from estimating population functions.

REFERENCES

Barr, A. J., Goodnight, J. H., Sall, J. P., and Helwig, J. T. A User's Guide to SAS 76. Raleigh, N.C.: SAS Institute, 1976.

_____. SAS Programmer's Guide. Raleigh, N.C.: SAS Institute, 1977.

David, F. N. Limiting Distributions Connected with Certain Methods of Sampling Human Populations. Stat. Res. Mem., 1938, 2.

Dixon, W. J. BMDP: Biomedical Computer Programs. Los Angeles: University of California Press, 1975.

Dunteman, G. H., Peng, S. S., and Holt, M. M. National Longitudinal Study of the High School Class of 1972 Composite Score Analysis: Ability Index, SES Index, Some Psychological and Educational Construct Scales. Research Triangle Park, N.C.: Research Triangle Institute, August 1974.

Durbin, J. Design of Multistage Surveys for the Estimation of Sampling Errors. Applied Statistics, 1967, 16, 152-164.

Feller, W. An Introduction to Probability Theory and Its Applications, Vol. II, Chapter 15. New York: Wiley, 1966.

Folsom, R. E. National Assessment Approach to Sampling Error Estimation. Sampling Error Monograph, Research Triangle Park, N.C.: Research Triangle Institute, 1974.

Folsom, R. E., Bayless, D. L., and Shah, B. V. Jackknifing for Variance Components in Complex Sample Survey Designs. Presented at the American Statistical Association Meetings at Fort Collins, Colorado, 1971.

Fuller, W. A. Regression Analysis for Sample Surveys. A report prepared for the U.S. Bureau of the Census on work conducted under the Joint Statistical Agreement, J.S.A. (Iowa State University), 1974.

Godambe, V. P. and Thompson, M. E. Bayes, Fiducial and Frequency Aspects of Statistical Inference in Regression Analysis in Survey-Sampling. Journal of Royal Statistical Society, B, 1971, 33, 361-390.

Gray, G. B. Components of Variance Model in Multi-Stage Stratified Samples. Survey Methodology, 1975, 1, 27-43.

Hájek, J. Limiting Distributions in Simple Random Sampling from a Finite Population. Pub. Math. Inst., Hungarian Acad. Sci., 1960, 5, 361-374.

Hidiroglou, M. A., Fuller, W. A., and Hickman, R. D. SUPER CARP, Survey Section, Statistical Laboratory, Iowa State University, 1975.

Holt, M. M. SURREGR: Standard Errors of Regression Coefficients from Sample Survey Data. Research Triangle Park, N.C.: Research Triangle Institute, 1977.

Horvitz, D. G. and Thompson, D. G. A Generalization of Sampling Without Replacement from a Finite Universe. Journal of American Statistical Association, 1952, 47, 663-685.

Institute for Social Research. OSIRIS III, Volume 1: System and Program Description. Ann Arbor, Mich.: The University of Michigan, 1973.

International Mathematical and Statistical Libraries. IMSL Library 1 Reference Manual. Houston: Author, 1975.

Kendall, M. G. and Stuart, A. The Advanced Theory of Statistics, Vol. I. London: Charles Griffin and Company, 1973, 231-323.

Kish, L. and Frankel, M. R. Inference from Complex Samples. Journal of Royal Statistical Society, B, 1974, 36, 1-37.

Kish, L., Frankel, M. R., and Van Eck, V. SEPP: Sampling Error Program Package, 1972.

Konijn, H. Regression Analysis in Sample Surveys. Journal of American Statistical Association, 1962, 57, 590-605.

Madow, W. G. On the Limiting Distribution of Estimates Based on Sample from Finite Universes. Inst. of Math. Stat., 1945, 19, 535-545.

McCarthy, P. J. Replication: An Approach to the Analysis of Data from Complex Surveys. Washington, D.C.: National Center for Health Statistics, Series 2, No. 14, 1966.

Moore, R. P., Shah, B. V., and Folsom, R. E. Efficiency Study of NLS Base-Year Design. Report on RTI Project No. 22U-884-3. Research Triangle Park, N.C.: Research Triangle Institute, 1974.

Nathan, G. and Holt, D. The Effect of Survey Design on Regression Analysis. Journal of Royal Statistical Society, B, 1980, 42, 377-386.

Nie, N. H., Hull, H., Jenkins, J. G., Steinbrenner, K., and Bent, D. H. SPSS: Statistical Package for the Social Sciences. New York: McGraw-Hill, 1975.

Public Health Service. Plan and Operation of the Health and Nutrition Examination Survey, DHEW publication No. (HSM) 73-1310, Series 1, Nos. 10a and 10b. Washington, D.C.: Government Printing Office, February 1973.

Quenouille, M. H. Notes on Bias in Estimation. Biometrika, 1956, 43, 353-360.

Rattenbury, J. and Eck, N. V. OSIRIS: Architecture and Design. Ann Arbor, Mich.: Institute for Social Research, The University of Michigan, 1973.

Royall, R. M. Linear Regression Models in Finite Population Sample Theory. Foundations of Statistical Inference, 1971, 259-274.

Shah, B. V. STDERR: Standard Errors Program for Sample Survey Data. Research Triangle Park, N.C.: Research Triangle Institute, 1974.

Shah, B. V. A Logic of Inference in Sample Survey Practice (unpublished manuscript). Research Triangle Park, N.C.: Research Triangle Institute, 1980.

Shah, B. V., Folsom, R. E., and Clayton, C. A. Efficiency Study of YEAR-03 In-School Design, Final Report 25U-796-2, Research Triangle Park, N.C.: Research Triangle Institute, 1973.

Shah, B. V., Holt, M. M., and Folsom, R. E. Inference About Regression Models from Sample Survey Data. Invited paper presented at the International Association of Survey Statisticians Third Annual Meeting, New Delhi, December 5-15, 1977.

Simmons, W. R. and Schnack, G. A. Development of the Design of the NCHS Hospital Discharge Survey. U.S. Department of Health, Education, and Welfare. Health Serv. and Mental Health Ad. Public Health Serv. Publ. No. 1000-Series 2-No. 39, 1970.

Tepping, B. J. The Estimation of Variance in Complex Surveys. Proceedings of the Social Statistics Section of the American Statistical Association, 1968, 11-18.

Tukey, J. W. Bias and Confidence in Not-Quite Large Samples. Abstract, Ann. Math. Statist., 1958, 29, 614.

U.S. Bureau of the Census. The Current Population Survey, A Report on Methodology. Technical Paper No. 7. Washington, D.C.: U.S. Government Printing Office, 1963.

Westat, Inc. Sample Design for the Selection of a Sample of Schools with Twelfth-Graders for a Longitudinal Study. Rockville, Md.: Author, 1972.

Wilkinson, J. H. and Reinsch, C. Linear Algebra. Edited by F. L. Bauer. Berlin: Springer-Verlag, 1971, 9-19.

Williams, R. L. The National Longitudinal Study: Estimated Third Follow-up Survey Design Effects and Associated Sampling Errors. Research Triangle Park, N.C.: Research Triangle Institute, October 1978.

Woodruff, R. S. A Simple Method for Approximating the Variance of a Complicated Estimate. Journal of American Statistical Society, 1971, 66, 411-414.

Woodruff, R. and Causey, B. D. Computerized Method for Approximating the Variance of a Complicated Estimator. Journal of American Statistical Society, 1976, 71, 315-321.

Appendix A

An Approximation to the Variance of
Regression Coefficients in Sample Surveys

Appendix A

An Approximation to the Variance of Regression Coefficients in Sample Surveys

In this appendix, the problem of estimating the variance of a vector of regression coefficients in a complex sample is solved by first finding a linear approximation to the estimator of the coefficients and then using this approximation to derive an approximation for the variance.

I. THE LINEARIZATION TECHNIQUE

The linearization technique employed in this paper is the Taylor Series expansion of the estimator. Tepping (1968) first used this approach with special reference to regression coefficients. Woodruff (1971) later elaborated it for a broad class of complex sample designs. In general, let $u = (u_1, u_2, u_3, \ldots, u_k)'$ be a vector of sample statistics and let $U = (U_1, U_2, \ldots, U_k)'$ represent a vector of simple population parameters such that $E[u] = U$. Let $f(U) = (f_1(U), \ldots, f_p(U))$ be a vector-valued function of U, which represents the p population parameters of interest. Assume that f(u) estimates f(U).

Now, f(u) is linearized by approximating it to its first-order Taylor Series expansion:

$$f(u) \cong f(U) + \sum_{i=1}^{k} (u_i - U_i) \frac{\partial f(U)}{\partial U_i}, \tag{A.1}$$

or

$$f(u) - f(U) \cong \sum_{i=1}^{k} (u_i - U_i) \frac{\partial f(U)}{\partial U_i}, \tag{A.2}$$

where

$$\frac{\partial f(U)}{\partial U_i} = \left[\frac{\partial f_1(U)}{\partial U_i}, \ldots, \frac{\partial f_p(U)}{\partial U_i} \right]' \tag{A.3}$$

Since $E[u_i - U_i] = 0$, it can be shown that $E[f(u) - f(U)] = 0$, to the order of approximation indicated. Consequently, the matrix form of the mean square error, where VAR indicates a variance-covariance matrix is

$$\text{VAR}[f(u) - f(U)] = E[\{f(u) - f(U)\}\{f(u) - f(U)\}'] \quad (A.4)$$

Using (A.2), (A.4) can be approximated by

$$\text{VAR}[f(u) - f(U)] = E[\{\sum_{i=1}^{k} (u_i - U_i) \frac{\partial f(U)}{\partial U_i}\}\{\sum_{j=1}^{k} (u_j - U_j) \frac{\partial f(U)}{\partial U_j}\}'] \quad (A.5)$$

Therefore,

$$\text{VAR}[f(u) - f(U)] = \sum_{i=1}^{k}\sum_{j=1}^{k} \left[\frac{\partial f(U)}{\partial U_i}\right]\left[\frac{\partial f(U)}{\partial U_j}\right]' \text{COV}(u_i, u_j) \quad (A.6)$$

where COV (.,.) is used to indicate the covariance of two entities.

If k is small, (A.6) is a convenient expression from which the variances of f(u) may be computed; however, if k is large (greater than 3), the formula becomes cumbersome. In this case, an alternative approach uses the actual numerical value of the sum of the k linearized portions of (A.2) so that the variance-covariance matrix of f(u) may be evaluated directly. Explicitly, define a new column vector, w, with p elements,

$$w = \sum_{i=1}^{k} (u_i - U_i) \frac{\partial f(U)}{\partial U_i} , \quad (A.7)$$

and observe that E[w] = 0. Now (A.5) can be expressed as

$$\text{VAR}[f(u) - f(U)] = E[ww'] = \text{VAR}[w] = \text{VAR}[z], \quad (A.8)$$

where

$$z = \sum_{i=1}^{k} u_i \frac{\partial f(U)}{\partial U_i} . \quad (A.9)$$

II. APPLICATION OF TAYLORIZED LINEARIZATION TO REGRESSION COEFFICIENTS

A realistic regression model may be defined using the notation from Section II of the report as

$$Y = XB + e. \quad (A.10)$$

Here e represents a vector of deviations from the linear prediction equation. Kish and Frankel's criterion minimizing the sum of the squared deviations over

the entire population yields a solution for B which is the familiar least squares solution to the normal equations:

$$B = (X'X)^{-1} X'Y \quad . \tag{A.11}$$

Now suppose that a sample, S, is drawn from the popultion and let the subscript i refer to any population number. If the units are selected with probability, P_i, then the unbiased Horvitz-Thompson estimators for $X'X$ and $X'Y$ are $x'x$ and $x'y$. (Lower case letters indicate sampling statistics.)

$$x'x = \sum_{i \varepsilon S} (X_i' X_i / P_i) \quad . \tag{A.12}$$

$$x'y = \sum_{i \varepsilon S} (X_i' Y_i / P_i) \quad . \tag{A.13}$$

The summations extend over units, i, belonging (ε) to the sample, S. The availability of unbiased estimates for $x'x$ and $x'y$ allows the estimation of B with

$$b = (x'x)^{-1}(x'y). \tag{A.14}$$

From (A.11) it can be seen that B is a function of $X'X$ and $X'Y$ while b is a function of $x'x$ and $x'y$ from (A.14). If it is assumed that there are p independent variables in the model, then $X'X$ and $x'x$ are pxp symmetric matrices. $X'Y$ and $x'y$ are pxl matrices. Let $(X'X)_{jj'}$, or $(x'x)_{jj'}$, represent the element of $X'X$ or $x'x$ in row j and column j'. Also let $(X'Y)_j$ or $(x'y)_j$ locate the row j element of $X'Y$ or $x'y$. Using the results presented in the previous section of this chapter, the Taylorized linearization of b can be written as

$$b \cong B + \sum_{j=1}^{p} [(x'y)_j - (X'Y)_j] \frac{\partial B}{\partial (x'Y)_j} \tag{A.15}$$

$$+ \sum_{j=1}^{p} \sum_{j'=j}^{p} [(x'x)_{jj'} - (X'X)_{jj'}] \frac{\partial B}{\partial (X'X)_{jj'}}$$

For regression coefficients, Tepping and Woodruff solved for the derivatives numerically. However, Folsom (1974) and Fuller (1974) independently developed an analytical expression for the derivatives, which simplifies the expression in (A.9). The remaining sections of this chapter follow Folsom's work. The partial derivatives are derived in Appendix B, and only the results are presented here.

For $j = 1, \ldots, p$ let d_j be the pxl column vector with a 1 in row j and zeros in all other rows. Also define $p(p+1)/2$ symmetric matrices, $D_{jj'}$, with dimension pxp and with zeros everywhere except in row j, column j' and row j', column j. These locations contain 1's.

From Appendix B we have

$$\frac{\partial B}{\partial (X'X)_{jj'}} = - (X'X)^{-1} D_{jj'} B , \tag{A.16}$$

and

$$\frac{\partial B}{\partial (X'Y)_j} = (X'X)^{-1} d_j . \tag{A.17}$$

Substituting (A.16) and (A.17) into (A.15) yields the approximation

$$b \cong B + (X'X)^{-1} \sum_{j=1}^{p} [(x'y)_j - (X'Y)_j] d_j \tag{A.18}$$

$$- (X'X)^{-1} \sum_{j=1}^{p} \sum_{j'=j}^{p} [(x'x)_{jj'} - (X'X)_{jj'}] D_{jj'} B .$$

Based on the definition of d_j and $D_{jj'}$, it can be seen that

$$\sum_{j=1}^{p} [(x'y)_j - (X'Y)_j] d_j = x'y - X'Y , \tag{A.19}$$

$$\sum_{j=1}^{p} \sum_{j'=j}^{p} [(x'x)_{jj'} - (X'X)_{jj'}] D_{jj'} = x'x - X'X . \tag{A.20}$$

Consequently,

$$b \cong B + (X'X)^{-1} [\{x'y - X'Y\} - \{x'x - X'X\}B] \tag{A.21}$$

$$= B + (X'X)^{-1} [x'y - (x'x)B + (X'X)B - X'Y] . \tag{A.22}$$

Using the fact that $(X'X)B = X'Y$,

$$b \cong B + (X'X)^{-1} [x'y - (x'x)B] , \tag{A.23}$$

or

$$b - B \cong (X'X)^{-1} [x'y - (x'x)B] . \tag{A.24}$$

III. APPLICATION TO THE STRATIFIED, TWO-STAGE SAMPLE DESIGN

For the purpose of this report, a stratified, two-stage sample design is assumed. In this type of design, the population has been divided into H strata by population or demographic characteristics. For stratum h (h=1,...,H) there are n(h) primary sampling units, PSUs. The actual observations are nested within each PSU, and there are n(hℓ) observations in PSU ℓ (ℓ=1,...,n(h)) within stratum h.

Referring to the first section of this chapter and remembering that the u_i are sample statistics, consider the case in which each u_i is a sum over sample observations of random values. (The regression problem represents such a case.) Let $u_i(h\ell j)$ indicate the observation for individual j (j=1,...,n(hℓ)) in PSU ℓ within stratum h.

For the stratified, two-stage sample design,

$$u_i = \sum_{h=1}^{H} \sum_{\ell=1}^{n(h)} \sum_{j=1}^{n(h\ell)} u_i(h\ell j) , \qquad (A.25)$$

and now (A.9) can be rewritten as the vector

$$z = \sum_{i=1}^{k} \left\{ \sum_{h=1}^{H} \sum_{\ell=1}^{n(h)} \sum_{j=1}^{n(h\ell)} u_i(h\ell j) \right\} \frac{\partial f(U)}{\partial U_i} . \qquad (A.26)$$

Rearranging the order of summation,

$$z = \sum_{h=1}^{H} \sum_{\ell=1}^{n(h)} \sum_{j=1}^{n(h\ell)} \left\{ \sum_{i=1}^{k} u_i(h\ell j) \frac{\partial f(U)}{\partial U_i} \right\} . \qquad (A.27)$$

Consequently, another vector, $z(h\ell)$, may be defined as

$$z(h\ell) = \sum_{j=1}^{n(h\ell)} \left\{ \sum_{i=1}^{k} u_i(h\ell j) \frac{\partial f(U)}{\partial U_i} \right\} , \qquad (A.28)$$

where

$$z = \sum_{h=1}^{H} \sum_{\ell=1}^{n(h)} z(h\ell) \qquad (A.29)$$

and

$$\text{VAR}\{z\} = \text{VAR}\left[\sum_{h=1}^{H}\sum_{\ell=1}^{n(h)} z(h\ell)\right], \tag{A.30}$$

for this sample design.

IV. THE GENERAL MEAN SQUARE ERROR FOR REGRESSION COEFFICIENTS

A biased, with-replacement approximation to the variance in (A.30) for a stratified two-stage sample design will be used. Gray (1975) states that the variance of a sample total from without-replacement sampling may be divided into a with-replacement variance component and a without-replacement covariance contribution at the first stage. By ignoring this covariance component, which is usually negative, a conservative approximation to the variance is obtained. It is usually assumed that this omitted finite population correction at the first stage is small and accounts for little of the total variance. This approximation for VAR[z] is

$$\sum_{h=1}^{H} n(h)S_z(h), \tag{A.31}$$

where

$$S_z(h) = \left[\sum_{\ell=1}^{n(h)} \{z(h\ell) - \bar{z}(h)\}\{z(h\ell) - \bar{z}(h)\}'\right]/\{n(h)-1\}, \tag{A.32}$$

and

$$\bar{z}(h) = \left[\sum_{\ell=1}^{n(h)} z(h\ell)\right]/n(h). \tag{A.33}$$

The actual specification of the approximation for the estimate of the variance of regression coefficients in a stratified two-stage sample design requires the definition of the row vector $X(h\ell j)$ as the X values for observation j in PSU ℓ and stratum h. Correspondingly, $Y(h\ell j)$ is the scalar response for a particular observation. From (A.12) and (A.13),

$$x'x = \sum_{h=1}^{H}\sum_{\ell=1}^{n(h)}\sum_{j=1}^{n(h\ell)} X'(h\ell j)X(h\ell j)/P(h\ell j) \tag{A.34}$$

and

$$x'y = \sum_{h=1}^{H} \sum_{\ell=1}^{n(h)} \sum_{j=1}^{n(h\ell)} X'(h\ell j)Y(h\ell j)/P(h\ell j). \qquad (A.35)$$

From the previous section of this appendix, $z(h\ell j)$ for the regression case can be defined as

$$z(h\ell j) = (X'X)^{-1}\,[X'(h\ell j)Y(h\ell j) - X'(h\ell j)X(h\ell j)B]/P(h\ell j). \qquad (A.36)$$

Now, the expression for $z(h\ell)$ can be written with one last level of approximation, which is imposed by substituting the estimates $(x'x)^{-1}$ and b, for $(X'X)^{-1}$ and B, respectively:

$$z(h\ell) = (x'x)^{-1} \sum_{j=1}^{n(h\ell)} [X'(h\ell j)\{Y(h\ell j) - X(h\ell j)b\}]/P(h\ell j). \qquad (A.37)$$

A convenient expression for (A.37) is obtained by defining the vector

$$r(h\ell j) = [X'(h\ell j)\{Y(h\ell j) - X(h\ell j)b\}]/P(h\ell j). \qquad (A.38)$$

Now (A.37) may be written as

$$z(h\ell) = (x'x)^{-1} \sum_{j=1}^{n(h\ell)} r(h\ell j). \qquad (A.39)$$

Appendix B

Derivation of the Partial Derivatives

Appendix B

Derivation of the Partial Derivatives

For the simplification of $\frac{\partial B}{\partial (X'Y)_j}$ and $\frac{\partial B}{\partial (X'X)_{jj'}}$ as found in (A.15), define for each row, r,

$$d_j(r) = \begin{cases} 1, & \text{if } j=r; \\ 0, & \text{otherwise;} \end{cases}$$

$$d_{jj'} = \begin{cases} 1, & \text{if } j = j'; \\ 0, & \text{otherwise.} \end{cases}$$

Also define $p(p+1)/2$ symmetric matrices, $D_{jj'}$, with dimension pxp and with zeros everywhere except in row j, column j' and row j', column j. These locations contain 1's. The element in row, r, and column, c, of $D_{jj'}$ can be written as

$$D_{jj'}(rc) = (1 - d_{jj'})d_j(r)d_{j'}(c) + d_{j'}(r)d_j(c).$$

Consider the partial derivatives of B with respect to each element in X'Y by taking the partials of both sides of the equality

$$(X'X)B = X'Y,$$

$$\frac{\partial (X'X)B}{\partial (X'Y)_j} = \frac{\partial (X'Y)}{\partial (X'Y)_j}, \quad j=1,2,\ldots,p,$$

$$\frac{\partial (X'X)B}{\partial (X'Y)_j} = d_j, \quad j=1,2,\ldots,p,$$

$$\frac{(X'X)B}{\partial (X'Y)_j} = d_j, \quad j=1,2,\ldots,p,$$

$$\frac{\partial B}{\partial (X'Y)_j} = (X'X)^{-1}d_j, \quad j=1,2,\ldots,p.$$

The derivation of the partials for X'X is more complicated. Again begin with the equality and observe that the right hand side is equal to zero after the derivatives with respect to each element of X'X are taken.

$$\frac{\partial(X'X)B}{\partial(X'X)_{jj'}} = \frac{\partial(X'Y)}{\partial(X'X)_{jj'}} = 0, \quad \begin{matrix} j=1,2,\ldots p, \\ j'=j,\ldots,p, \end{matrix}$$

$$\frac{\partial(X'X)B}{\partial(X'X)_{jj'}} = \left[\frac{\partial(X'X)}{\partial(X'X)_{jj'}}\right] B + (X'X)\left[\frac{\partial B}{\partial(X'X)_{jj'}}\right] = 0,$$

$$(X'X)\frac{\partial B}{\partial(X'X)_{jj'}} = \left[\frac{\partial(X'X)}{\partial(X'X)_{jj'}}\right] B = -D_{jj'}B.$$

Consequently,

$$\frac{\partial B}{\partial(X'X)_{jj'}} = -(X'X)^{-1} D_{jj'}B, \quad \begin{matrix} j=1,2,\ldots,p, \\ j'=j,\ldots,p. \end{matrix}$$

Appendix C

The Survey Regression Procedure

Appendix C

The Survey Regression Procedure

This appendix provides a brief description of a flexible program developed by the authors for estimating regression parameters and associated standard errors from data arising from survey samples. The procedure employed is based on a Taylor Series Linearization approach, described in Appendix A. The program, entitled SURREGR, has been incorporated into an existing statistical analysis package, Statistical Analysis System (SAS); a Users Manual for the program is provided in Appendix D.

I. GENERAL

One of the most difficult tasks in providing a new, flexible, statistical computer program is in convincing programmers who know little about statistics and statisticians and social scientists who know little about programming to use it. Experience has shown that statistical programs that stand alone with their own specialized control cards are avoided whenever possible. For maximum utility, these programs need to operate within a system which takes care of interfacing with the user; however, it is extremely time-consuming to design and implement such a system. Therefore, it was decided that the survey regression program would be written to run under an existing statistical system.

Several statistical packages, BMDP, SPSS, OSIRIS, and SAS, were reviewed, and, among these, it was determined that SAS possessed the best data management capabilities. The particular advantages of SAS were as follows:

a) the ability to extensively manipulate the input data,
b) the immediate availability of other types of statistical analyses,
c) free format of procedure information statements,
d) comprehensive error checking for data and procedure information statements,
e) procedure output as a SAS data set which is available for further analysis, and
f) the dynamic allocation of core which enables flexible programming within system core and time limitations.

Furthermore, programming details and technical assistance were readily available within the local computing facility, the Triangle Universities Computation Center (TUCC). Consequently, it was decided that the survey regression program would be written as an SAS procedure. The SAS documentation was provided by Barr, Goodnight, Sall, and Helwig (1976 and 1977).

II. COMPUTATIONAL PROCEDURES

The survey regression procedure, SURREGR, has five main functions:

a) interpretation of user input,
b) accumulation of sums of squares and cross products,
c) a solution for the regression coefficients,
d) general mean square errors, and
e) tests of hypothesis.

The approach taken for each function is discussed in the following subsections.

A. Interpretation of User Input

This function is controlled by the language module, which is independent of the computational part of the program and is responsible for the parsing of the SAS language statements. Although the language module must be written in IBM 360 assembler language, SAS macros are provided for the standard parsing of the variable lists, options, and parameters. The philosophy for the parsing of the model statement is borrowed from the SAS general linear models procedure, GLM. The GLM language model was modified to allow for multiple model statements within one execution of SURREGR and to permit effects and interactions formed by categorical dependent variables. For all categorical variables that are declared as effects or interactions in the model statement, SURREGR generates the required number of binary (0,1) variables, q_i. These dummy variables are defined as

$$q_i = \begin{cases} 1, \text{ if an observation has a particular value for that variable;} \\ 0, \text{ otherwise.} \end{cases} \qquad \text{(C.1)}$$

Only after all information statements are parsed without error will SAS execute the computational part of the program.

B. Accumulation of Sums of Squares and Cross Products

The X'WX and X'WY matrices are computed as the second main function of the procedure. These matrices are accumulated as sums of squares and cross products of variables over all observations. In other words, the actual X, Y, and W matrices are never formed. However, the W matrix can be represented as a square diagonal matrix with the number of rows and columns equal to the number of observations and with equal diagonal element containing an observation's weight. The X and Y matrices are defined in Section II of the report with one exception--there may be more than one column allocated for an effect (which may be a classificatory variable.) Interactions among continuous and classificatory variables are permitted.

C. A Solution for the Regression Coefficients

To compute such a solution, the inverse of X'WX must be found. The Cholesky decomposition technique described by Wilkinson and Reinsch (1971) is used to compute a standard matrix inverse unless X'WX is singular. In this case, a generalized inverse is computed. This inverse, A^{-1}, for a matrix, A, must satisfy the following conditions:

$$A = AA^{-1}A \quad , \tag{C.2}$$

and

$$A^{-1} = A^{-1}AA^{-1} \quad . \tag{C.3}$$

A check for the numerical accuracy of equations (C.2) and (C.3) is provided since some ill-conditioned matrices may be subject to large numerical errors. Each term on the right-hand side of the equations is evaluated and compared with the corresponding element on the left-hand side. The maximum difference found between any two elements of either comparison is reported to the user. If any deviation exceeds a set tolerance, the user is given a warning message; however, the program will continue. Subsequently, the regression coefficients are computed by the formula in (A.14) given in Appendix A.

D. General Mean Square Errors

This computation requires that the file be reread and that the Taylorized deviations defined in Appendix A, equation (A.38), be computed. These deviations

may be rewritten in the notation of this appendix as

$$r(h\ell j) = [X'(h\ell j)\{Yh\ell j) - X(h\ell j)B\}]W(h\ell j), \tag{C.4}$$

where $r(h\ell j)$ is a column vector, and $W(h\ell j) = 1/P(h\ell j)$; other notations are defined in Appendix A and are not repeated here. The following sums and sums of squares and cross-products are computed

$$r(h\ell) = \sum_{j=1}^{n(h\ell)} r(h\ell j) , \tag{C.5}$$

$$rr'(h) = \sum_{\ell=1}^{n(h)} r(h\ell)r'(h\ell) , \tag{C.6}$$

$$r(h) = \sum_{\ell=1}^{n(h)} r(h\ell)r'(h\ell) , \tag{C.7}$$

The variance-covariance matrix is then accumulated over strata and adjusted by $(X'WX)^{-1}$ following these accumulations to yield the variance covariance matrix S_b^2. Specifically,

$$S_b^2 = (X'WX)^{-1} \left[\sum_{h=1}^{H} [\{n(h)rr'(h) - r(h)r'(h)\}/\{n(h)-1\}] \right] (X'WX)^{-1} . \tag{C.8}$$

E. Tests of Hypothesis

The last major function of the program is to compute the tests of hypothesis first for the entire model and then for each effect. The null hypothesis for any of these tests may be written as

$$H_o: B_m = B_{m+1} = \ldots = B_n = 0, (\text{for } n \geq m), \tag{C.9}$$

against the alternative hypothesis

$$H_1: B_k \neq 0 \text{ (for some } k, m \leq k \leq n). \tag{C.10}$$

For a particular hypothesis, the program determines its rank, d, and a dxp matrix, C, such that the given hypothesis is in the form $CB = 0$. The value of d is n-m+1, if all of the parameters, B_m, B_{m+1}, ..., B_n, are estimable. Otherwise, the value of d is less than n-m+1.

If the parameters were normally distributed, the test statistic would be a likelihood ratio criterion which would have an approximate F distribution for large samples. If S_B^2 is the variance-covariance matrix of B with degrees

of freedom, e, equal to the number of PSU's minus the number of strata, then the test statistic from Folsom (1974),

$$F_{d,e} = \left(\frac{e-d+1}{ed}\right) (CB)'(C\, S_B^2\, C')^{-1}(CB), \qquad (C.11)$$

is an approximate F with d and e degrees of freedom under the null hypothesis.

III. DESIGN FEATURES

The SURREGR procedure is designed to produce a regression anaysis for sample survey data. To achieve this end, a number of unique features have been incorporated into the program. The following attributes place SURREGR in a class apart from the standard regression packages of BMDP, SPSS, and SAS:

a) SURREGR accounts for the correlation between observations due to the sample design.

b) There is no program limit to the number of models which may be specified in one procedure.

c) Effects and interactions are allowable in independent and dependent variables.

d) Standard tests of hypotheses are provided, and, in the case of a non-full-rank hypothesis, a test of the estimable subhypotheses is made.

e) Checks are made to establish the condition of $(X'WX)^{-1}$.

f) SURREGR has the ability to select multiple random samples from a data file which is considered to be a finite population. This permits empirical evaluation of the performance of the statistics and tests generated by the program.

Appendix D

User's Manual for the SURREGR Procedure

Appendix D

User's Manual for the SURREGR Procedure

SURREGR is a procedure which provides a means of producing appropriate tests of hypotheses for regression models in sample survey situations. The procedure offers many useful options and operates in three modes, which differ only in the method by which the variance-covariance matrix of the regression coefficients is calculated. SURREGR was developed principally to handle regression analysis for sample survey data; hence, the default mode of the procedure will incorporate a stratified multistage sampling design into the variance-covariance computation. Another mode of the procedure relies on the ordinary least squares estimate for the variance-covariance matrix. Finally, a weight may be used for a weighted ordinary least squares analysis.

The Procedure SURREGR Statement

PROC SURREGR options and parameters;

The options and parameters for the PROC SURREGR statement are grouped by function.

FILE OUTPUT

DATAOUT (abbreviated DOUT)

This option produces a SAS file which contains for each model the regression coefficients, the variance-covariance matrix, the F test values, and their associated degrees of freedom. A new record is generated for each different value of a dependent effect. The data record structure and output variable names and descriptions follow:

MODEL	Model number
DVAR	Dependent variable number
NCELL	Number of columns of the X matrix
NTESTS	Number of F test values
CHECK	This variable equals zero if the X'X inverse matrix is acceptable.

B001-B_ _ _The regression coefficients (beta values). Each variable name for a regression coefficient starts with a B and ends with a three-digit number with leading zeros, which is the column of the X matrix to which the regression coefficient corresponds. B001, for example, represents the intercept value if an intercept was included in the model.

V001-V_ _ _The variance-covariance matrix of the regression coefficients. The matrix is output in lower triangular form by rows. The variable name starts with a V and ends with a three-digit number with leading zeros, which is the position of the variable in the lower triangular matrix.

F001-F_ _ _The F test values from the tests of hypothesis for the entire model (F001) and for each independent effect (F002-F_ _ _).

D001-D_ _ _The degrees of freedom associated with each F test value.

It is important to realize that the specification of the output data set cannot be made with the standard SAS two-level format. A separate parameter is needed for each level.

DDNAME=_ _ _(abbreviated DDN)

This parameter is used with the DATAOUT option to specify the DDNAME in a JCL statement which describes the OS data set for the output file. If DDNAME is omitted a temporary file will be used.

DSNAME=_ _ _(abbreviated DSN)

DSNAME is used with the DATAOUT option. It is a six-character name for the output data set. If DSNAME is omitted the name, DUMMYM, will be generated by the procedure. Since each different model produces a different data set, a two-character suffix to the six-character data set name is added by the procedure to identify the model number. These two characters range from 01 through the number of models.

RESIDUAL

This option allows for output to an SAS data set of the unweighted predicted and residual values associated with each level of each dependent

effect for each model. There is one output record for each observation in the input file. The output variables are:

M01PRD01-M01PRD_ _The predicted values. The variable name begins with the letter M, followed by a two-digit number, the letters PRD, and a two-digit number with leading zeros for the dependent effect level. For example, the predicted value for the second continuous dependent variable in the fourth model is M04PRD02.

M01RSD01-M01RSD_ _The residuals. (Description the same as for predicted values.)

OUT=_ _ _ This parameter is associated with the RESIDUAL option. It provides the procedure with a standard one- or two-level SAS data set name. If it is omitted the next WORK data set will be used.

PRINTED OUTPUT

The hypothesis testing results and the checks on the inverse of X'X are printed by default.

NOPRINT This option suppresses all printed output.

BETA BETA prints a solution to the normal equations and the variance-covariance matrix for that solution. It should be noted that singularities in the X matrix produce corresponding zeros in the regression coefficients and in the variance-covariance matrix. There is no reparameterization.

XPX This option prints the X'X matrix and its inverse.

MODES OF OPERATION

Computation of the variance-covariance matrix using Taylorized deviations and a sampling structure is default.

OLS OLS requests ordinary least squares analysis.

WLS WLS requests weighted ordinary least squares analysis.

TAYWLS TAYWLS will compute WLS and then will repeat the analysis using the sampling structure and Taylorized deviations.

FILE INPUT

DATA=_ _ _This parameter specifies a standard one- or two-level SAS data set name to be used by the procedure as the input data. If DATA are omitted, the current SAS data set will be used.

PROGRAM CONTROL

MISSPSU This option is for Taylorized deviations and is only needed when no more than one PSU (primary sampling unit) in a stratum has valid data. A divisor used in computing the variance-covariance matrix of the regression coefficients is corrected from the number of PSUs in a stratum with valid data minus 1 to the total number of PSUs in a stratum minus 1.

TOL=_ _ _The absolute tolerance used to compute all relative tolerances in the program is set at 10^{-8} unless this parameter is assigned a different value.

PLACES=_ _ _The number of digits used for all matrix printing is set to 8 unless this parameter is assigned a different value.

PROCEDURE INFORMATION STATEMENTS

Model Statement

MODEL dependent effects = independent effects/list of options;

The MODEL statements allow the user to list one or multiple dependent effects with any number of independent effects. An effect may be a single variable or a main effect, or it may be composed of a group of variables. When there is more than one variable in an effect, each variable must be joined to the next with either a * indicating crossed variables or a ( ) indicating a nesting structure. An effect may contain continuous or discrete variables, but only discrete variables may be nested. Variables which are combined into one effect must be listed with the crossed and then the nested groupings. Only one level of nesting is allowed.

Examples of correctly formed effects:

A*B	A crossed with B.
A(B)	A nested within B.
A*B(C)	A crossed with C nested within B.
A(B C)	A nested within B crossed with C.

Note that an * is not to be used before the (or between B and C.

Examples of incorrectly formed effects:

A*(B)	The * is not allowed.
A(B)*C	Crossing must be specified before nesting.
X1-X10 or A*(X1-X5)	All variables must be individually listed. The "-" option of SAS is not valid in the MODEL statement.
NOINT	Only one option is available for a particular MODEL statement. Unless NOINT is specified, SURREGR will assume an intercept for the model.

CLASSES STATEMENT (abbreviated CLASS)

CLASSES list of variables; in order for a variable to be treated as discrete, it must be in the CLASS statement. CLASSES A1-A4 is a valid CLASS statement.

PSU STATEMENT

PSU variable name; PSU gives the name of the variable containing a numerical primary sampling unit indicator.

STRATUM STATEMENT (abbreviated STR)

STRATUM variable name; STRATUM gives the name of the variable containing a numerical representation for each stratum. Remember that the data set must be sorted by PSU within stratum for a Taylorized deviation computation of the variance-covariance matrix.

WEIGHT STATEMENT (abbreviated WT)

WEIGHT variable name; WEIGHT gives the name of the sampling weight variable.

LEVELS STATEMENT

LEVELS list of numbers separated by blanks; a level is the number of values available for a particular discrete variable. That variable must be coded from 1 through the maximum value available. There must be a level specified for each variable listed in the CLASSES statement and the levels must be ordered exactly as the variables in the CLASSES statement. If there are variables in the

CLASSES statement which all have the same number of levels, then the notation can be shortened. Four consecutive variables with two levels each may be written as:

LEVELS 4*2; or LEVELS 2 2 2 2.

Certain modes of SURREGR require different procedure information statements:

Statement	Taylorized Deviations	OLS	WLS	TAYWLS
MODEL	required	required	required	required
CLASSES	optional	optional	optional	optional
PSU	required	irrelevant	irrelevant	required
STRATUM	required	irrelevant	irrelevant	required
WEIGHT	required	not allowed	required	required
LEVELS	required with the classes statement			

COMPUTATIONAL METHODS AND NOTES

The X Matrix

The X matrix is a matrix with a row for each observation. The number of columns is the sum of the number of locations needed to hold each effect plus one column for the intercept if necessary. A continuous main effect or an effect with continuous variables crossed together requires only one column. A discrete main effect requires columns equal to the number of levels for that variable. When discrete variables are crossed or nested, the number of columns is equal to the product of the levels for each variable. The values of the effect are located within the program as well as for output by varying the value of the last variable most rapidly. If an effect is defined at A*B*C where A has a levels, B has b levels, and C has c levels, then the actual location of an observation A=x, B=y, and C=z within the a*b*c available locations is (x-1)*b*c + (y-1)*c + z. Note that X'X is accumulated once for all dependent variables in a model. In order to have different treatments for different dependent variables, a separate MODEL statement must be used for each dependent variable.

CHECKS ON THE INVERSE

The sums of squares and cross products matrix for the independent effects in a model statement, $X'X$, is inverted as a part of the least squares procedure. The inverse, $(X'X)^{-1}$, is a generalized inverse. There is a check provided on the condition of the inverse. Each element of $X'X$ is compared with that of $X'X\ (X'X)^{-1}\ X'X$ and then each element of $(X'X)^{-1}$ is compared with that of $(X'X)^{-1}\ X'X\ (X'X)^{-1}$. A relative deviation equal to (the check value minus the actual value) divided by the actual value is compared with the program's set tolerance times 100. If any deviation exceeds the set tolerance, the user is given a warning message.

THE VARIANCE-COVARIANCE MATRIX OF THE REGRESSION COEFFICIENTS

If no option relating to the variance-covariance matrix is specified, a between-PSU (primary sampling unit), within-stratum, generalized mean square error (GMSE) is computed. This GMSE is derived for the regression problem using the technique of Taylorized linearization yielding a Taylorized deviation which is incorporated in the computations.

For the OLS option, the variance-covariance matrix is $(X'X)^{-1}\ \hat{\sigma}^2$.

$$\hat{\sigma}^2 = (Y'Y - b'X'Y) / (N-r)$$

where Y is a vector of all observations for one dependent effect, b is a vector of regression coefficients for that dependent effect, N is the number of observations, and r is the rank of X.

For the WLS option, the variance-covariance matrix has the same formula as for OLS except that each product of dependent and independent effects - observation has been multiplied once by that observation's weight.

HYPOTHESIS TESTING

The last major function of the program is to compute the tests of hypothesis first for the entire model and then for each effect. These tests exclude the intercept. The null hypothesis for any of these tests may be written as

$$H_0 : B_m = B_{m+1} = \ldots = B_n = 0 (\text{for } n \geq m),$$

against the alternative hypothesis

$$H_1 : B_k \neq 0 \text{ (for some } k,\ m \leq k \leq n).$$

For a particular hypothesis, the program determines its rank, d, of the estimable subspace of the hypothesis and a dxp matrix, C, such that the parameters CB are estimable and rank C is d. The value of d is n - m + 1, if all of the parameters, B_m, B_{m+1}, ... B_n, are estimable. Otherwise, the value of d is less than n - m + 1.

If the parameters were normally distributed, the test statistic would be a likelihood ratio criterion which would have an approximate F distribution for large samples. If S_B^2 is the variance-covariance matrix of B with degrees of freedom, e, equal to the number of PSUs minus the number of strata minus the rank of X'X, then the test statistic,

$$F_{d,e} = \left(\frac{e-d+1}{ed}\right) (CB)'(C\, S_B^2\, C')^{-1} (CB)$$

is an approximate F with d and e degrees of freedom under the null hypothesis. For OLS, e is equal to the number of observations minus the rank of X'X.